Features of the Pilot Study and Desired Features of a Post-FACDQ Pilot Study

Report Out From Technical Work Group
06/22/06

Introduction

The Federal Advisory Committee on Detection and Quantitation Approaches in Clean Water Act Programs (FACDQ) decided that conducting a pilot study of several procedures would provide information that would help the committee develop its final recommendations. At its March 2006 meeting the committee tasked the Technical Work Group and a Design Team (one representative from each caucus) with developing a pilot study design with which EPA could begin the laboratory contracting process this spring. The Design Team and the Technical Work Group completed a design that EPA is using to solicit bids for participation in the pilot. It is anticipated that the committee will ratify the design at its July 13-14, 2006 meeting so that the pilot study may begin in August.

As the Technical Work Group members developed the pilot study design they agreed that there were many positive items that will come from the pilot. They also understood that there are questions the pilot will not answer that might be addressed in a post-FACDQ pilot study. The Technical Work Group acknowledged the need to consider this document dynamic in that much can be learned about what the pilot study will accomplish as the pilot study moves forward and the data are analyzed. Additionally, the list of features suggested in this document should not be interpreted as an exhaustive list of all desired features of a post-FACDQ pilot study.

This document provides an overview of what the pilot will and will not accomplish, tradeoffs made in developing the pilot design, and recommended elements for a post-FACDQ pilot study. The document begins with brief descriptions of what the pilot does/does not do, and the tradeoffs in putting this design together, followed by descriptions in response to thirteen statements in the FACDQs document titled “What we need a procedure to do.”

Overview of the Pilot

Summary of What the Pilot Does:

The pilot will test three detection and three quantitation procedures, far more than what was envisioned early in the committee process. Initially, the committee considered pilot testing one procedure for detection and one for quantitation.

The pilot will use eight laboratories per each of five analytical methods tested, and as many as 12 different concentrations (based on whether the procedure is a regression design or a single lab design). Because not every laboratory is expected to analyze samples with all five methods, the

number of labs participating in the pilot likely will be two to three times more than the minimum of eight.

Samples will be tested over approximately 15 working days, and will therefore capture some within-laboratory variability. The study will calculate and confirm limits for each procedure being evaluated.

Details of what the pilot will do are described later in this document under sections entitled, "What the Pilot Study Does."

Summary of What the Pilot Will Not Do:
It should be noted that while the pilot will address several of the needs identified in the FACDQ's "What We Need a Procedure to Do," it will not address everything. It is important that FACDQ expectations of the pilot are in line with what the pilot will actually do.

There are several procedure components that will not be tested in the pilot, primarily those items related to long-term or ongoing verification steps included in some of the piloted procedures. The following are some specific examples of issues the pilot study will not test. The pilot study will not:

- Include long-term verification steps
- Include determination and verification of MQOs using long-term analysis results
- Prepare confirmation spikes at the exact value of the determined limits
- Confirm if a procedure yields a biased-high L_C
- Include additional sources of intra-lab variability between instruments and analysts
- Test real world matrices
- Test all analytical methods
- Evaluate intermittent contamination
- Adequately evaluate long-term variability

For more detail on these points, refer to the additional sections in this document entitled "Desired Features of a Post-FACDQ Pilot Study", as well as in the comments sections.

Tradeoffs

The Design Team and the Technical Work Group agreed early that while there is substantial funding available for this effort, this funding is limited. They understood that there is a limited time period for conducting the pilot and analyzing the data. The pilot study was designed with these conditions in mind. Tradeoffs were considered and evaluated on what could realistically be accomplished while still providing value to the committee.

Both groups recognized that the earliest month the actual pilot testing could begin was August 2006. While this is late in the FACDQ process, there are specific reasons why this later start date became necessary. Those are:

- The pilot was designed concurrently with committee decisions on policy issues such as uses of detection and quantitation, which are still ongoing.
- The pilot design needed a committee decision on measurement quality objectives (MQOs) that the detection and/or quantitation limits would target, and what the committee needs a procedure to do, both of which were decided at the March 2006 committee meeting.
- Some of the procedures were modified to meet committee needs, such as MQOs and the elements in the "What we need a procedure to do" document.
- Before pilot design efforts went too far, it was necessary to conduct a substantial review of existing data and studies that would inform MQO decisions and frame the pilot study.

The Design Team and Technical Work Group identified the following tradeoffs in the current pilot study design:

- The timeframe for collecting laboratory data is a maximum of 45 calendar days, however labs are required to test over a 15 working day period to collect data, which meant that the pilot study will not provide ongoing, quarterly, or annual verification of results as specified in some of the procedures.
- The pilot study will not confirm results by subsequent spiking at the calculated limits, though this could be done in a post-FACDQ pilot.
- The pilot study will only use reagent water matrices, though real world matrices could be tested in a post-FACDQ pilot.
- Although the pilot study will evaluate five analytical methodologies, there are others that may be desirable for testing in a post-FACDQ pilot.
- The pilot study will not capture some sources of intra-lab variability such as variability between analysts and between instruments, though a post-FACDQ pilot could do this.

Responding to the FACDQs "What We Need Procedures to Do" Document

The following section compares what the pilot will do versus what a post-FACDQ pilot could address using thirteen statements in the FACDQ document as a template for responses.

The procedure(s) should:

1. **provide an explicit estimate of bias at L_Q for limits that must be verifiable by labs at those limits.**

   To be evaluated by:

   a. reviewing procedure(s) and specifically identifying the quantitative limit for bias at L_Q that is tested in the pilot study.

b. requiring labs to analyze samples (spikes, blind or otherwise as appropriate) and comparing observed bias to that cited by the procedure(s).

2. **provide an explicit estimate of precision at L_Q for limits that must be verifiable by labs at those limits.**

   To be evaluated by:

   a. reviewing procedure(s) and specifically identifying the quantitative limit for precision at L_Q that is tested in the pilot study.
   b. requiring labs to analyze samples (spikes, blind or otherwise as appropriate) and comparing observed precision to that cited by the procedure(s).

What the Pilot Study Does

The pilot will provide numeric estimates of both bias and precision at L_Q for the analytes tested. The pilot study will generate estimates of L_Q for three different procedures, and will also generate estimates of precision and bias at each of these calculated limits. These short term precision and bias estimates will then be compared to the target MQOs to determine if they were achieved.

Desired Features of a Post-FACDQ Pilot Study

For single laboratory procedures that incorporate long term estimates, a post-FACDQ pilot study should also derive long term estimates (n=20 to100 over 6 to12 months), if one of these single lab procedures is considered in a post-FACDQ pilot. The post-FACDQ pilot study labs could analyze additional samples for limit calculations and confirmation/verification.

More spiked sample results would be needed to be able to perform statistical significance tests comparing estimates of bias or precision at a given quantitation limit to the target MQO values. Since it is a primary objective of the FACDQ to evaluate the long term precision and bias performance of the laboratories against the estimated L_Q (for both the single and regression based procedures), the post-FACDQ pilot study should incorporate an additional 20-40 blind spikes at the estimated L_Q over a 6-12 month period.

Comments

- The estimates of precision generated in the single laboratory pilot study design (ACIL procedure) will not reflect the long-term precision of the laboratory and/or method. The reason for this is that the observed precision during the study will consist of 20 observations collected simultaneously (same period of time over which the estimates themselves will be produced) over a three to six consecutive week period. The estimates of bias generated in the pilot study will be variable due to the small number of observations and may not reflect some factors that occur rarely during a long time period, but may affect the method performance.
- While this limitation applies equally to the regression based procedures (Hubaux-Vos, LC-MRL, and IDE/IQE procedures), the regression based procedures do not

incorporate long-term precision and bias estimates. However, it is still the objective of the FACDQ to evaluate the long term robustness of these estimates through the use of long term confirmation/verification data.

- Limited analyses of existing data suggest that procedures based entirely on single concentration, low level spike precision may adequately address long period variability. If the data gathering spans a 21 day (3 week) period, these data are limited both in terms of how many laboratories were evaluated and in the methods that were evaluated. In addition, these analyses were based on results at a single concentration, and the relationship between time and variability may differ for measurements made over a range of concentrations.
- Any confirmation pilot should confirm the final recommended procedure(s) by estimating the parameter and then preparing spikes at that exact level for subsequent confirmation. The proposed confirmation approach will use spikes close to the estimate and then rely on interpolation or modeling to determine if the MQOs were met. Thus, the strength of the confirmation is only as good as the validity of the interpolation or modeling process.

3. **provide an explicit false positive rate for L_C**

   To be evaluated by:

   a. reviewing procedure(s) and specifically identifying the false positive error rate predicted for each limit that is tested in the pilot study.
   b. comparing the false positive rate of lab blanks at the levels of L_C to those predicted by the procedure(s).

What the Pilot Study Does

It is difficult for any pilot to estimate the committee's target one percent false positive rate with high precision without several hundred blanks. If the observed false positive error rate is, for example 5%, then fewer blanks (only about 20-50) would be required to make this determination. Nonetheless, the pilot will document false positive results (i.e., those results exceeding the estimates of Lc generated by three different procedures and comparing the rate of false positive results to the target of less than or equal to1% false positives). Depending on the amount of blank data submitted by the participating laboratories, the calculation and confirmation of the estimates of L_C may encompass as much as 6 months of variability.

Desired Features of a Post-FACDQ Pilot Study

It is difficult for any pilot to determine a one percent false positive rate with high precision without several hundred blanks. If the observed false positive error rate is slightly higher (for example 2%) then fewer blanks (only about 50-100) would be required to make this determination.

Additional blank sample results would be needed to be able to perform statistical significance tests comparing observed false positive rates at a given detection limit to the target MQO value.

A post-FACDQ pilot should attempt to improve information related to this MQO by evaluating the long-term and on-going false positive procedures for estimating L_C for all single lab procedures under consideration. Any short-term estimates of L_C estimated by the procedure(s) that are tested in the post-FACDQ pilot study should also be verified for the single laboratory and/or regression based designs. To accomplish this, at least 100 blanks should be analyzed over a 6 to 12 month period to test whether the L_C is set too low or too high. If the procedure generates a biased low L_C (i.e., generates false positive rates > 1%) then fewer blank analyses would be required.

Comments

- The ACIL single laboratory procedure includes specific directions to evaluate the on-going long-term performance of the estimates they generate. This includes both parametric and non-parametric evaluations and adjustments to the estimates over time.
    a. This is particularly important because most of the short- and long-term estimates are based on assumptions of normality, which may be incorrect the majority of the time.
    b. The initial short- and long-term estimates must have larger safety factors built in because they are based on smaller amounts of data. As a result these estimates are larger.
    c. The initial estimates are also less accurate and less representative of routine laboratory performance. A period of six months to a year would be necessary to properly evaluate the on-going performance of a laboratory.
- Because L_C estimates are being developed concurrently with the verification of those estimates, the result will be more favorable confirmatory results than if the confirmation process were performed over a separate time period. Although, for uncensored methods where existing blank data are used to generate the initial estimates of L_C, subsequent verification of false positive error rates will be possible.
- Due to experimental design and cost limitations, it is difficult to confirm that a procedure yields a biased-high L_C (i.e., produces estimates that yield too low of a false positive rate).
    a. To determine whether or not a procedure yields L_C's with false positive error rates that are significantly below 1%, with any degree of confidence, would require an enormous amount of data (a minimum of 100 replicates). This is not a criticism of the pilot design but a practical reality that may not have a practical solution, but which could be implemented in the subsequent EPA confirmation work.

4. **provide an explicit false negative rate for L_C for the true value at L_D and L_Q that must be observed in labs at L_C for the estimated values of L_D and L_Q.**

    To be evaluated by:

a. reviewing procedure(s) and specifically identifying the false negative error rate predicted for L_D/L_Q that is tested in the pilot study.
b. comparing the false negative rate of results obtained by analyzing samples spiked at L_D/L_Q concentration to those predicted by the procedure(s).

What the Pilot Study Does

It is difficult for any pilot to estimate the committee's target one percent false negative rate with high precision without several hundred samples spiked at the L_D or L_Q. If the observed false negative error rate is for example 5%, then fewer spiked samples (only about 20-50) would be required to make this determination. The pilot will document false negative results – those results less then the estimated value of L_C based on making the detection decision at L_C (and/or based on instrument signal for censored methods) for true concentrations equal to L_D (as generated by the IDE procedure) and L_Q (as generated by the ACIL, LC-MRL, and IQE procedures), and compare the rates to the target MQO of 1%.

Desired Features of a Post-FACDQ Pilot Study

It is difficult for any pilot to determine a one percent false negative rate with high precision without several hundred low level spikes. If the observed false positive error rate is slightly higher (for example 2%) then fewer spikes (only about 50-100) would be required to make this determination.

More spiked sample results would be needed to be able to perform statistical significance tests comparing observed false negative rates at a given quantitation limit to the target MQO value. A post-FACDQ pilot study should evaluate the long-term and on-going procedures for estimating false negatives at L_D, and L_Q for all single lab procedures under consideration. Any short-term estimates of L_D and/or L_Q estimated by the procedure(s) that are tested in the pilot study should also be verified for the single laboratory and/or regression based designs. To accomplish this at least 100 spikes at false negatives at L_D, or L_Q could be analyzed over a 6 to 12 month period to test whether the L_D or L_Q are set too high or too low. If the procedure generates a biased low L_D or L_Q (i.e. generates false negative rates > 1%) then fewer spike sample analyses will be required.

Comments

- To actually test whether an estimate of L_D or L_Q achieves the target false negative error rate will require a large number of low level spikes (Minimum of 100).

5. **provide that qualitative identification criteria defined in the analytical method are met at the determined detection and quantitation limits.**
   To be evaluated by:
   a. requiring that all method qualitative identification criteria be satisfied in order for detection to occur.
   b. requiring modification of L_Q or L_D if all spikes at L_Q or L_D are not detected.

What the Pilot Study Does

The pilot study requires the labs to follow the procedures and analytical methods as they are written, so qualitative identification criteria will be determined as defined in the analytical method. However, the pilot study requires the labs to provide indications of where they did not follow the criteria. The study coordinator will determine if all method qualitative identifications were actually met.

Desired Features of a Post-FACDQ Pilot Study

If it is found in the pilot study that laboratories did not apply method qualitative criteria as specified in the method or that the laboratories interpreted the qualitative identification criteria differently, clarified wording could be developed for the method(s) and the procedures retested.

Comments

- Allowing labs to have the flexibility to set their own qualitative identification criteria or interpret the criteria set forth in the method differently, is truly representative of current routine laboratory operation.
    a. This will not provide information regarding actual adherence to the criteria set forth in the methods.
    b. Although this may be a realistic reflection of the latitudes that labs have taken in performing the promulgated methods, we will not know if the performance of the L_C or L_D procedures being evaluated is at fault for failing to estimate the MQOs, or if it is due to different labs using different and disparate identification criteria.

6. **adequately represent variability in lab performance.**

    To be evaluated by determining whether the procedures:

    a. use data to calculate limits that are collected over enough time to capture variability in performance relative to MQOs.
    b. recalculate limits at a frequency that captures variability in performance relative to MQOs.
    c. incorporate variability due to the use of multiple instruments per lab.
    d. incorporate variability due to use of multiple analysts per lab.
    e. incorporate variability occurring across laboratories (not for single lab procedure).
    f. Adjust or account for recovery.
    g. provide recommendations or limit choices for outlier tests.
    h. address varying numbers of different concentrations (spikes) that can be used between laboratories (may only apply to multi/inter lab procedures).
    i. address varying numbers of replicates per concentration (spike) that can be used between laboratories (may only apply to multi/inter lab procedures).
    j. address varying combinations of concentrations (spikes) that can be used between laboratories (may only apply to multi/inter lab procedures).

k. adequately accommodate different models of instruments used per analyte and technology to calculate limits.

What the Pilot Study Does

The pilot study will capture many of the items listed above including:

e. incorporate variability occurring across laboratories for the ASTM procedures (IDE/IQE).
f. address varying analyte recovery.
h. address varying numbers of different concentrations (spikes) that can be used between laboratories for pilot design, not for final procedure
i. address varying numbers of replicates per concentration (spike) that can be used between laboratories for pilot design, not for final procedure (high priority)
j. address varying combinations of concentrations (spikes) that can be used between laboratories for pilot design, not for final procedure. (high priority)
k. adequately accommodate different models of instruments used per analyte and technology to calculate limits.

Desired Features of a Post-FACDQ Pilot Study

The six week time period over which the study will occur may not adequately represent variability in performance. A post-FACDQ pilot study should be done over enough time (e.g., 6-12 months) and with enough labs and instruments that these variables will be able to be evaluated more fully, especially:

a. use data to calculate limits that are collected over enough time to capture variability in performance relative to MQOs.
c. incorporate variability due to the use of multiple instruments per lab.
d. incorporate variability due to use of multiple analysts per lab.

Comments

- While analyses of existing data suggest that procedures based entirely on single concentration, low level spike precision may adequately address long period variability when data are gathered over the pilot's 21 day (3 week) period, the pilot data are collected from eight laboratories per analytical method, and five analytical methods. However, because not every laboratory is expected to analyze samples with all five methods, the number of labs participating in the pilot likely will be two to three times more than the minimum of eight per method.
- Because the on-going verification part of the ACIL procedure is not evaluated in the pilot study, the process of recalculating limits over time cannot be evaluated.
- Because we are using the same labs and same instruments to develop the estimates of L_C, L_D and L_Q and for the confirmation, the pilot study will not evaluate how well the procedures estimate the performance of analytical methods being analyzed by the full population of labs.

a. This is only relevant to either the multi-lab procedure, if its intent is to characterize the method's performance by qualified laboratories, and/or the interlab procedures.
b. The ideal confirmation would be to use one set of labs to develop the estimates and a separate set of labs to test how reliable the estimates were.

Note: This laboratory and instrument variability may be better assessed, if enough volunteer labs are available.

- Outlier testing may be evaluated for the inter-laboratory procedures, but there will be a limited opportunity to test this on the single laboratory procedure, because of the small amount of data (N=20) generated during the pilot study.

7. be capable of calculating limits using matrices other than lab reagent grade water.

To be evaluated by:

a. reviewing procedures and determining that there is nothing precluding the use of matrices other than reagent grade water to calculate limits.
b. reviewing procedures to determine if they incorporate steps to verify when limits adopted for an analytical method can or cannot be met in a matrix other than lab reagent grade water.
c. reviewing procedures to determine if they provide instructions on preparing an analyte-free matrix that approximates the matrix in question.

What the Pilot Study Does

The pilot does not test matrices other than reagent water, but the study coordinator can evaluate the procedures to see if they are written in a manner to accommodate other matrices and identify matrix problems.

Desired Features of a Post-FACDQ Pilot Study

A limited number of real world matrices and the procedures for evaluating real world matrices could be tested in a post-FACDQ pilot.

8. use only data that results from test methods conducted in their entirety.

To be evaluated by determining whether the procedure(s):

a. require that samples used to calculate detection and quantitation limits undergo all routine steps outlined in an analytical method as specified in the laboratory's SOP (prep method, extraction, etc.)
b. reviewing procedures to determine if they incorporate steps to verify when limits adopted for an analytical method can or cannot be met when a sequence of non-routine steps are used.

What the Pilot Study Does

The pilot does require the use of all steps of a method, so this will be tested for the five analytical methods that will be used in the pilot.

Desired Features of a Post-FACDQ Pilot Study

Same as pilot study, however a few additional methods or preparative steps could be added if some gaps need to be filled.

9. **explicitly adjust or account for situations where method blanks always return a non-zero result/response.**

   To be evaluated by:

   a. reviewing the procedure(s) and determining if they include a procedure to address occasions where method blanks always return a non-zero result.
   b. reviewing the procedure(s) and determining if they require calculation of statistics regarding non-zero results/responses.
   c. reviewing the procedure(s) and determining if they mathematically adjust limits for non-zero results/responses.

What the Pilot Study Does

All procedures being evaluated in the pilot study address methods where the results/responses are not equal to zero.

Desired Features of a Post-FACDQ Pilot Study

Same as pilot study.

10. **explicitly adjust or account for situations where method blanks are intermittently contaminated.**

    To be evaluated by:

    a. reviewing the procedure(s) and determining if they define intermittent contamination and provide explicit instructions to deal with this situation.
    b. reviewing the procedure(s) and determining if they mathematically adjust limits for non-zero results/responses

What the Pilot Study Does

None of the procedures being tested in the pilot address intermittent contamination.

Desired Features of a Post-FACDQ Pilot Study

A new or revised procedure that considers intermittent blank contamination could be evaluated in a post-FACDQ pilot as long as it ran long enough for intermittent contamination to be observed.

Comments

- The pilot study as planned will not evaluate the impact of either of the following practices on the estimate of L_C.

a. If the laboratories reject the batch and exclude the blank in the L_C estimate when there is intermittent contamination under normal practice.
b. If the laboratories qualify the batch and then include the blank with intermittent contamination in the L_C estimate under normal practice. However, it is problematic that a lab would eliminate all data from a batch based on a contaminated blank, yet assume that blank is representative of routine performance by retaining the result for limit calculation purposes.

11. **be clearly written with enough detail so that most users can understand and implement them.**

To be evaluated by:

a. asking users to interpret data prior to our after-procedure calculations are carried out. Examples include What is the resulting detection limit?, What is the resulting quantitation limit? and What is the blank bias?
b. asking users questions about the procedure characteristics, using the matrix as a point of reference. Examples include Do the procedures address recovery?, How often is a limit calculated by the user?, and How often is data generated to calculate limits for a given procedure?
c. asking users to perform calculations or run software and interpret results.
d. asking users to select spikes for given circumstances.
e. reviewing the procedure(s) and determining which ones minimize the amount of data required to calculate analytical limits beyond that normally generated by analytical methods.
f. determining that the procedure(s) do not require skills of users in addition to those that are normally required by laboratories.

What the Pilot Study Does

The pilot will ask participating laboratories to opine on whether the ACIL and Hubaux-Vos/LC-MRL procedures are clearly written with enough detail so that most users can understand and implement them. The clarity and completeness of the IDE/IQE pair will not be evaluated by the participating laboratories.

Desired Features of a Post-FACDQ Pilot Study

The ultimate FACDQ recommendations for detection and quantitation procedures will further narrow the procedures, so a post-FACDQ pilot study should evaluate the clarity and completeness of any procedure ultimately recommended by the FACDQ.

Comments

- The pilot study as planned will evaluate the following procedures.

a. The IDE/IQE has a "regression" study design team that should evaluate the procedures for overall clarity/detail, and the labs will be evaluated for selection of spike levels.
b. The LCMRL and Hubaux-Vos procedures - both individual labs and the study team can evaluate the procedure for overall clarity and detail.
c. The ACIL single laboratory procedure where individual labs can evaluate the procedures for overall clarity and detail.

- While the labs will review all procedure steps, they may not actually test all verification steps.

12. **be cost effective.**

To be evaluated by:

a. reviewing the procedure(s) and determining which ones minimize the amount of data required to calculate analytical limits beyond that normally generated by analytical methods.
b. determining whether the procedure(s) require the purchase of software or equipment in addition to that which is normally required by laboratories.
c. determining that the procedure(s) do not require skills of users in addition to those that are normally required by laboratories.

What the Pilot Study Does

The pilot study will address this somewhat through the narrative responses from the labs for those methods tested.

Desired Features of a Post-FACDQ Pilot Study

Only the methods being evaluated directly by multiple laboratories in the pilot can be evaluated for cost-effectiveness, so a confirming pilot should look at expanding the number of methods tested. Also, any follow-up determinations of limits as required by a procedure (ongoing, quarterly, annual, etc.) should be evaluated for costs in a confirming pilot. A post-FACDQ pilot study could evaluate the recommended detection/quantitation procedure(s) with a broader range of labs, analysts, methods, and matrices.

13. **be applicable to all users and test methods.**

To be evaluated by:

a. testing procedures against objectives 1 – 13 among a representative sample of labs (states, EPA, commercial, municipal, small, medium and large, etc.) considering cost and contracting restraints.
b. testing procedures against objectives 1 – 13 among a representative sample of analytical test methods (different technologies and analytes)

What the Pilot Study Does

The pilot is testing the three pairs of detection/quantitation procedures among a minimum of eight labs per analytical method; five analytical methods covering a range of chemical, detectors, and analytical approaches (GC, LC, IC, ICP, etc.) chemistries are being tested.

Desired Features of a Post-FACDQ Pilot Study

While the pilot study is intentionally designed to evaluate a representative group of test methods and as many different types of users as possible, it will not be possible to evaluate every test method and every user type. A post-FACDQ pilot study could evaluate the recommended detection/quantitation procedure(s) with a broader range of labs, analysts, methods, and matrices.